Application of discrete equations to forecasting (New Developments in the Research of Integrable Systems : Continuous, Discrete, Ultra-discrete)

AUTHOR(S):

Satoh, Daisuke

CITATION:

Satoh, Daisuke. Application of discrete equations to forecasting (New Developments in the Research of Integrable Systems : Continuous, Discrete, Ultra-discrete). 数理解析研究所講究録 2003, 1302: 116-136

ISSUE DATE:

2003-02

URL:

http://hdl.handle.net/2433/42743

数理解析研究所講究録 1302 巻 2003 年 116-136

離散方程式の予測への応用

(Application of discrete equations to forecasting)

NTT サービスインテグレーション基盤研究所　佐藤大輔 (Daisuke SATOH)
NTT Service Integration Laboratories

1 Introduction

From the end of the 1990's, discrete integrable equations have been appearing in many fields, e.g., algorithms and traffic flow [17, 18, 19, 20, 34]. Discrete integrable systems are expected to be applied to engineering.

Forecasting is important in engineering. The making of decisions in various industries is heavily reliant on forecasting. Forecasting is the dominant factor in decisions on how many finished products should be made, how much stock should be prepared and so on. In the past there has been a tendency for forecasts to be unduly optimistic. This has produced some serious problems. Therefore, the accuracy of forecasting is of great importance.

Growth curve models are used for forecasting in many fields, e.g., ecology [3, 25, 35], agriculture [26], life sciences [13], marketing [1, 12], and software reliability growth models (SRGMs) [14, 22, 36]. To forecast the ceiling, we estimate parameters of the differential equation which provides the growth curve model. The differential equations which are used generally have exact solutions. In the conventional method, the differential equation's parameters are estimated by using an ordinary forward or central difference equation as its approximation. Generally, the ordinary forward or central difference equation does not have an exact solution. Therefore, the difference equation does not conserve the properties of the differential equation.

Although a growth curve model has practical applications, one generally known point is that the model does not provide accurate parameter estimates using the data available during the early phases of the process being forecast. The conventional method is only capable of providing accurate estimates of parameters at the end of the phase. For forecasting to be of practical value, accurate estimates must be obtained early in the phase.

In this paper, the application of discrete integrable equations to forecasting is discussed. We focus on discrete integrable analogues of the logistic equation, the Gompertz equation, and the Riccati equation for forecasting in two fields: marketing and SRGM. The remainder of this paper is organized as follows. From Sect. 2 to Sect. 4 and in Sect. 6, we consider the forecasting of numbers of software faults or software failures through an SRGM.

In Sect. 2, we describe discrete analogues of the logistic curve model [30], which has been observed in the testing of software systems [23, 27]. The model is described by either of two difference equations, which were proposed by Morishita [15] and Hirota [5, 6], respectively. We will see that both models yield accurate parameter estimates, even when there is only a small amount of input data from actual software testing.

Although the logistic curve model is one of the S-shaped SRGMs, S-shaped software reliability growth for actual projects is often more closely described by the Gompertz curve than by the logistic curve [2, 8, 21]. In Sect. 3, we consider the Gompertz curve as an SRGM. Firstly, we propose a discrete Gompertz equation [28] that has an exact solution. We will see that the proposed model provides accurate estimates of parameters, enabling prediction early in the testing phase of when the software will be fit for release.

There is a further problem for software engi-

neers and managers: they have had little guidance as to which models are likely to be best for a particular application. In Sect. 4, we propose a criterion [31], together with a discrete SRGM, for determining the absolute worth of a model.

In Sect. 5, we consider the Bass model, which is the main impetus underlying behind the recent diffusion research in marketing. The author has previously proposed a discrete form of the Bass model [29]. This model provides more accurate estimates of parameters than is possible with the conventional Bass model. Furthermore, parameter estimation of the discrete Bass model overcomes the three shortcomings of parameter estimation by the conventional (continuous) Bass model: the time-interval bias, standard error, and multicolinearity.

The proposed models yield accurate estimates of parameters, even from small amounts of input data. These models, however, are deterministic equations, so they do not yield distributions of the estimates. In Sect. 6, we propose a discrete stochastic logistic equation that have an exact solution and describe an SRGM that is based on this equation. This model yields distributions of an estimate along with the estimates themselves.

Finally, in Sect. 7, we summarize the results of this paper.

2 Logistic curve model

2.1 Conventional logistic curve model

The logistic curve model is described as

$$\frac{dL(t)}{dt} = \frac{\alpha}{k} L(t)(k - L(t)), \tag{1}$$

where $L(t)$ is the cumulative number of software failures occurred up to testing time t and α and k are constant parameters to be estimated through regression analysis.

A solution of Eq. (1) is given by

$$L(t) = \frac{k}{1 + m\exp(-\alpha t)}, \tag{2}$$

where $k > 0$, $m > 0$, and $\alpha > 0$. The parameter k represents the total number of potential software failures occurring over an infinitely long duration or the initial number of faults inherent in the software system.

2.1.1 Conventional parameter estimation 1

Regression analysis is generally used to estimate total numbers of potential software failures, although there is a further conventional method of estimation, which is described in Sect. 2.1.2.

We take the following regression equation:

$$Y_n = A + BL_n, \tag{3}$$

where

$$t_n = n\delta, \tag{4}$$

$$L_n = L(n\delta), \text{ and} \tag{5}$$

$$Y_n = \frac{\frac{L_{n+1} - L_{n-1}}{2\delta}}{L_n}. \tag{6}$$

Here, δ is a constant difference interval.

Given regression coefficients $\hat{A}$ and $\hat{B}$, where $\hat{A}$ means the value of A as estimated through regression analysis, estimates of the parameters k, α, and m can be obtained as

$$\hat{k} = \frac{\hat{A}}{\hat{B}}, \tag{7}$$

$$\hat{\alpha} = \hat{A}, \text{ and} \tag{8}$$

$$\hat{m} = \frac{\sum_{n=1}^{N} (\hat{k} - L_n)}{\sum_{n=1}^{N} (L_n \exp(-\hat{\alpha} t_n))}. \tag{9}$$

These estimates depend on the difference interval δ, because Eq. (6) depends on δ.

The accuracy of estimates thus derived is said to be poor when there are only a few data points. For accuracy, we require data points up to at least one point after the point of inflection ($t_n = \frac{\log m}{\alpha}, L_n = \frac{k}{2}$). We can judge whether the obtained data includes the point of inflection by checking whether or not $\frac{\bar{k}}{2} < L_n$ is satisfied, where $\bar{k}$ is predicted empirically or statistically.

For the estimates of parameters to be reliable, the following condition must be satisfied:

$$w\bar{k} < L_n, \tag{10}$$

where $\bar{k}$ is predicted empirically or statistically and w is a constant parameter, the value of which is empirically chosen from the range 0.6 to 0.8 [14].

2.1.2 Conventional parameter estimation 2

Another conventional method of estimation is based on a modified exponential curve model [24]. This model is described as

$$y = c + ba^t. \tag{11}$$

We rewrite the logistic curve model as

$$\frac{1}{L(t)} = \frac{1}{k} + \frac{m}{k}\exp(-\alpha t). \tag{12}$$

This equation is in the form of the modified exponential curve model.

When it is possible to place a model in this form, parameters a, b, and c are estimated through the following method of estimation. At first, we divide the data set into three subsets, each of which has the same number of data points. If the number of data points is not a multiple of three, we discard the first one or two points. Then we sum up the data in each subset. Finally, parameters a, b, and c are obtained as,

$$a = \left(\frac{S_3 - S_2}{S_2 - S_1}\right), \tag{13}$$

$$b = (S_2 - S_1)\frac{a-1}{(a^n-1)^2}, \tag{14}$$

$$c = \frac{1}{n}\left\{S_1 + (S_1 - S_2)\frac{1}{a^n - 1}\right\}, \tag{15}$$

where S_1, S_2, and S_3 represent the summations of all elements of the first, second, and third subsets of the data, respectively, and n represents the number of data points in each of the subsets.

We then obtain estimates of the parameters k, α, m by using these estimators:

$$k = \frac{1}{c}, \tag{16}$$

$$\alpha = -\log a, \text{ and} \tag{17}$$

$$m = \frac{b}{c}. \tag{18}$$

2.2 Discrete logistic curve models

Two discrete analogues of the differential equation (1) for the logistic curve model have already been proposed. We propose a regression equation that is appropriate for the estimation of parameter for use with these equations.

2.2.1 Discrete logistic curve model with Morishita's equation

Morishita [15] proposed the following equation as a discrete form of Eq. (1):

$$L_{n+1} - L_n = \delta\frac{\alpha}{k}L_{n+1}(k - L_n). \tag{19}$$

It has an exact solution:

$$L_n = \frac{k}{1 + m(1-\delta\alpha)^{\frac{t_n}{\delta}}}, \tag{20}$$

where $t_n = n\delta$.

Let $\alpha_c = \alpha$ in Eq. (2), and let $\alpha_{dm} = \alpha$ in Eq. (20). Comparing Eqs. (2) and (20), we get

$$\alpha_c = -\frac{1}{\delta}\log(1 - \delta\alpha_{dm}). \tag{21}$$

To derive the regression equation for the parameters k, α_{dm}, and m, we rewrite Eq. (19) as

$$Y_n = A + BL_{n+1}, \tag{22}$$

where

$$Y_n = \frac{L_{n+1}}{L_n}, \tag{23}$$

$$A = \frac{1}{1 - \delta\alpha_{dm}}, \tag{24}$$

$$B = -\frac{\delta\alpha_{dm}}{k(1 - \delta\alpha_{dm})}, \text{ and} \tag{25}$$

$$t_n = n\delta. \tag{26}$$

Parameters k, α, and m are estimated by

$$\hat{k} = \frac{1 - \hat{A}}{\hat{B}}, \tag{27}$$

$$\delta\hat{\alpha}_{dm} = 1 - \frac{1}{\hat{A}}, \tag{28}$$

$$\hat{m} = \frac{\sum_{n=1}^{N}(\hat{k} - L_n)}{\sum_{n=1}^{N}(L_n(1 - \delta\hat{\alpha}_{dm})^n)}, \tag{29}$$

where $\hat{A}$ and $\hat{B}$ are the estimates of parameters A and B, respectively.

Y_n in Eq. (22) is independent of the difference interval δ, because δ is not used in this equation. The estimates of $\hat{k}$, $\delta\hat{\alpha}_{dm}$, and $\hat{m}$ are the same whatever value of δ we choose.

2.2.2 Discrete logistic curve model with Hirota's equation

Hirota [5, 6] discretized Eq. (1) as

$$L_{n+1} - L_n = \delta\frac{\alpha}{k}L_n(k - L_{n+1}). \quad (30)$$

He gave this exact solution:

$$L_n = \frac{k}{1 + m(\frac{1}{1+\delta\alpha})^{\frac{t_n}{\delta}}}, \quad (31)$$

where $t_n = n\delta$.

Let $\alpha_{dh} = \alpha$ in Eq. (31). Comparing Eqs. (2) and (31), we get

$$\alpha_c = \frac{1}{\delta}\log(1 + \delta\alpha_{dh}). \quad (32)$$

To derive the regression equation for parameters k, α, and m, we rewrite Eq. (30) as

$$Y_n = A + BL_{n+1}, \quad (33)$$

where

$$Y_n = \frac{L_{n+1}}{L_n}, \quad (34)$$

$$A = \delta\alpha_{dh} + 1, \quad (35)$$

$$B = -\frac{\delta\alpha_{dh}}{k}, \text{ and} \quad (36)$$

$$t_n = n\delta. \quad (37)$$

The estimates of parameters k, α, and m are given as by

$$\hat{k} = \frac{1 - \hat{A}}{\hat{B}}, \quad (38)$$

$$\delta\hat{\alpha}_{dh} = \hat{A} - 1, \quad (39)$$

$$\hat{m} = \frac{\sum_{n=1}^{N}(\hat{k} - L_n)}{\sum_{n=1}^{N}\left(L_n(\frac{1}{1+\delta\hat{\alpha}_{dh}})^n\right)}, \quad (40)$$

where $\hat{A}$ and $\hat{B}$ are the estimates of parameters A and B, respectively.

Y_n in Eq. (33) is independent of the difference interval δ because δ is not used in Eq. (33). The same estimates of $\hat{k}$, $\delta\hat{\alpha}_{dh}$, and $\hat{m}$ are obtained, whatever value of δ we choose.

The regression equation (33) is the same as Eq. (22). Moreover, the same estimate of k is given by both equations. Though the estimate of α depends on discrete equations, both discrete equations yield the same estimate of α_c. The same estimate of m is obtained because

$$1 - \delta\alpha_{dm} = \frac{1}{1 + \delta\alpha_{dh}} = \exp(\alpha_c) = \frac{1}{\hat{A}}. \quad (41)$$

Therefore, the models with Morishita's and Hirota's equations both give the same value of L_n.

2.3 Parameter estimation in the logistic curve models

We compared the accuracy of parameter estimation for the first conventional logistic curve model and the discrete logistic curve models. To compare only the accuracy of parameter estimation, we evaluated the performance of the parameter estimates when the data represented an exact solution of the logistic equation. We did not consider the second conventional method of parameter estimation, as described in Sect. 2.1.2, because it inherently reproduced the target values of the parameters when given data that were an exact solution of the logistic equation.

We prepared data that represented exact solutions of the logistic equation for a set of periods ($t = 0$ to 21). We set $k = 100$, $\alpha = 0.8$, and $m = 999$ as the target values. This data was inflected at the point where $t^* = 8.63$ and $L(t^*) = 50$. In our evaluation, we set the difference interval to 1. We analyzed four sets of this data: the three data sets that covered an data up to (i) the ceiling ($t = 0, 1, \ldots, 21$), (ii) just after the point of inflection ($t = 0, 1, \ldots, 9$), (iii) just before the point of inflection ($t = 0, 1, \ldots, 8$), and (iv) the set of the first three data points($t = 0, 1, 2$).

The results of our comparison are shown in Table 1. Since we used an exact solution as the input data, an accurate method of estimation should reproduce the parameters that generated this solution. Table 1 shows that the proposed models estimated k correctly, even when the data set only consisted of the first

Table 1: Estimated parameter k.

	Conventional model 1	Proposed models
i	99.229	100
ii	79.782	100
iii	72.168	100
iv	60.166	100

four points. The conventional model had lower accuracy, despite the use of exact solutions to the differential equation as data values. As was earlier stated, the conventional model is generally known to provide poor estimates of the parameters in the situation represented by data sets (iii) and (iv), i.e., when the data set does not include the data points around the point of inflection. Empirical studies have shown that stable and robust estimates of the parameters of SRGMs, such as the logistic curve model, cannot be obtained without using data points that cover the point of inflection and satisfy Eq. (10), i.e., $w\bar{k} < L_n$. Even when data set (ii) was used, the conventional method provided estimates the parameter values that were neither stable nor robust, even though set (ii) both includes the point of inflection and satisfies Eq. (10).

We evaluated the discrete logistic curve models on an actual data set to show that they are more appropriate to use than the conventional model. The data were debugging data for an item of software. We evaluated the parameter estimates given both with all data and with only that data available early in the testing phase. In our evaluation, we set δ equal to 1. Figure 1 shows results for the 'all data' case; we see that having all of the data leads to all three models fitting the actual data very well.

Moreover, the discrete logistic curve models have the important advantage of providing accurate parameter estimates early in the testing phase as well as at the end of the testing phase. Therefore, the accuracy of an SRGM's parameter estimates early in the testing phase is an important aspect of its utility as an estimator. The accuracy of parameter k is especially important, because this parameter indicates the

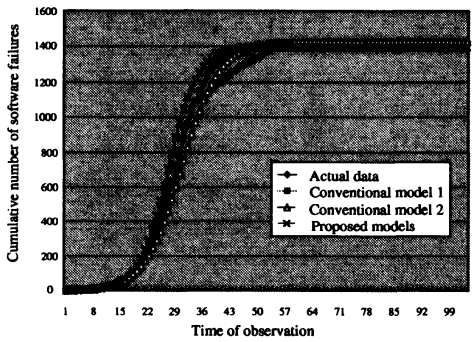

Figure 1: Comparison the three models with actual data.

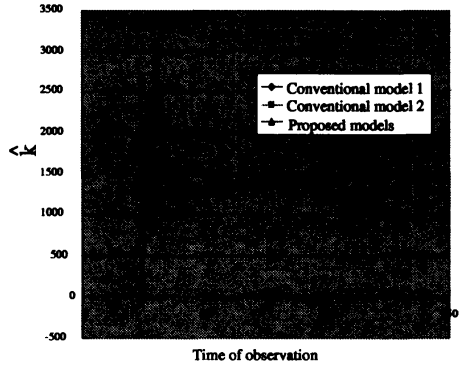

Figure 2: Parameter estimates of k.

potential fault content of the software system.

To compare the conventional and proposed models in terms of this property, we estimated values of k from increasing number of data points, starting with a small amount of data from the earlier part of the testing phase. As is shown in Fig. 2, the values of k estimated by the proposed models were stabler than those estimated by the conventional models. Since the proposed models provide more accurate estimates of parameter values from small amounts of data gathered early in the testing phase, they provide better estimates of total numbers of potential software failures, k, early in the testing phase.

3 Gompertz curve model

3.1 Conventional Gompertz curve model

The Gompertz curve model is described as

$$\frac{dG(t)}{dt} = G(t)(\log a)(\log b)b^t \tag{42}$$

or

$$\frac{dG(t)}{dt} = (\log b)G(t)\log\frac{G(t)}{k}, \tag{43}$$

where $G(t)$ is the cumulative number of software failures detected up to a testing time t. By integrating either equation and assuming that $G(0) = ka$, $G(t)$ can be written as

$$G(t) = ka^{b^t} \quad (k > 0,\ 0 < a < 1,\ 0 < b < 1), \tag{44}$$

where a, b, and k are parameters whose constant values are estimated by using regression analysis. Parameter k represents the total number of software failures with the potential to occur over an infinitely long period or the initial fault content in the software system:

$$G(t) \to k \quad (t \to \infty). \tag{45}$$

3.1.1 Conventional parameter estimation 1

Regression analysis is generally used to estimate total numbers of potential software failures, although we also have the conventional method of estimation that is shown in Sect. 3.1.2.

The following regression equation is obtained:

$$Y_n = A + Bn, \tag{46}$$

where

$$Y_n = \log\left(\frac{G_{n+1} - G_{n-1}}{2\delta G_n}\right), \tag{47}$$

$$A = \log((\log a)(\log b)), \text{ and} \tag{48}$$

$$B = \delta \log b. \tag{49}$$

Given regression coefficients $\hat{A}$ and $\hat{B}$, where $\hat{A}$ means the parameter A as estimated through regression analysis, we have these estimators for parameters a, b, and k:

$$\hat{a} = \exp\left(\frac{\delta \exp\hat{A}}{\hat{B}}\right), \tag{50}$$

$$\hat{b} = \exp\left(\frac{\hat{B}}{\delta}\right), \text{ and} \tag{51}$$

$$\hat{k} = \frac{\sum_{n=1}^{N} G_n}{\sum_{n=1}^{N} \hat{a}^{\hat{b}^{\delta n}}}. \tag{52}$$

These estimates depend on the difference interval δ, because Y_n in Eq. (46) depends on δ. We can choose any value as δ. Therefore, the estimates are entirely dependent on the specific value of δ.

The accuracy of these estimates is poor when there are only a few data points. We need data up to at least one point after the point of inflection to get accurate estimates. A further condition must be satisfied:

$$w\bar{k} < G_n, \tag{53}$$

where $\bar{k}$ and w are the same as in the case of the logistic curve model.

3.1.2 Conventional parameter estimation 2

The other method of estimation is the same as that of Sect. 2.1.2. The Gompertz curve model can be rewritten as

$$\log G(t) = \log k + (\log a)b^t. \tag{54}$$

This equation is in the form of the modified exponential curve model.

Parameters k, a, and b are estimated as

$$k = \exp\left[\frac{1}{n}\left\{S_1 + \frac{S_1 - S_2}{a^n - 1}\right\}\right], \tag{55}$$

$$a = \exp\left(\frac{(S_2 - S_1)(a-1)}{(a^n - 1)^2}\right), \text{and} \tag{56}$$

$$b = \frac{S_3 - S_2}{S_2 - S_1}, \tag{57}$$

where S_1, S_2, and S_3 represent the summations of the first, second, and third sets as defined in Sect. 2.1.2 of data, respectively, and n represents the number of data points in each set.

3.2 Discrete Gompertz equation

We propose a discrete analogue of Eq. (42) for the Gompertz curve model:

$$G_{n+1} = G_n \left(\frac{G_n}{k}\right)^{\delta \log b}. \tag{58}$$

The exact solution of this equation is

$$G_n = k a^{(1+\delta \log b)^n}, \tag{59}$$

where $k > 0, 0 < a < 1$, and $\frac{1}{e} < b^\delta < 1$. Equation (59) satisfies Eq. (45) given any δ:

$$G_n \to k \quad (n \to \infty). \tag{60}$$

3.3 Discrete Gompertz curve model

From Eq. (58), the regression equation is obtained:

$$Y_n = A + B \log G_n, \tag{61}$$

where

$$\begin{aligned} Y_n &= \log G_{n+1} - \log G_n, && (62)\\ A &= -\delta (\log b)(\log k), \text{ and} && (63)\\ B &= \delta \log b. && (64) \end{aligned}$$

Using Eq. (61), we can estimate parameters k, a, and b:

$$\begin{aligned} \hat{k} &= \exp\left(-\frac{\hat{A}}{\hat{B}}\right), && (65)\\ \hat{a} &= \exp\left(\frac{\sum_{n=1}^{N} \log \frac{G_n}{\hat{k}}}{\sum_{n=1}^{N} (1+\delta \log \hat{b})^n}\right), \text{and} && (66)\\ \hat{b} &= \exp\left(\frac{\hat{B}}{\delta}\right), && (67) \end{aligned}$$

where $\hat{a}$, $\hat{b}$, and $\hat{k}$ are the estimated values of a, b, and k, and $\hat{A}$ and $\hat{B}$ are the estimated values of A and B in Eq. (61).

Y_n in Eq. (61) is independent of difference interval δ because δ is not used in Eq. (61). Hence, the estimates of $\hat{k}$, $\hat{a}$, and $\delta \log \hat{b}$ are the same, regardless of our choice of value for δ. Therefore, Eq. (59) is determined uniquely for any value of δ.

We evaluated the performance in parameter estimation by the discrete Gompertz model when given data that was an exact solution of the Gompertz equation. We did this by comparing the accuracy of the parameters estimated by the conventional Gompertz curve model 1 and by its discrete form.

To restrict our comparison to the accuracy of parameter estimation. We used parameter values of $k = 100$, $a = 0.01$, and $b = 0.5$ as target values in preparing data that represented exact solutions of Eq. (43) for a set of periods ($t = 0$ to 25). This data was inflected at the point where $t^* = 2.20325$ and $G(t^*) = 36.7879441$.

We analyzed three sets of this data: they covered the data up to (i) the ceiling ($t = 0, 1, \cdots, 25$), (ii) just after the point of inflection ($t = 0, 1, 2, 3$), and (iii) just before the point of inflection ($t = 0, 1, 2$).

The result of the comparisons is shown in Table 2. The value of k as estimated by using the proposed discrete model matched the target value for all three data sets.

Since we used an exact solution as the input data, an accurate method of estimation should reproduce the parameters that generated this solution. Table 2 shows that the proposed model estimated k correctly, even when the data did not include the point of inflection. The accuracy of conventional model 1 was poor, despite the use of exact solutions to the differential equation as data values. The conventional model is generally known to provide poor estimates of the parameters in the situation represented by data set (iii), i.e., when the data set does not include the data points around the point of inflection.

Even when data set (ii) was used, the estimates of parameters provided by the conventional method were neither stable nor robust, even though this set does include the point of inflection and satisfies Eq. (53).

Table 2: Estimated parameter k.

	Conventional model 1	Proposed model
i	99.631	100
ii	78.159	100
iii	46.529	100

3.4 Model evaluation with actual data

We used actual data in evaluating the discrete Gompertz curve model. We evaluated the parameter estimates both with all data and with the data available early in the test phase. We used the same data as had been used by Mitsuhashi [14].

The time scale δ is not used in the regression equation of the proposed discrete model, but is used in the equation of the first conventional method. Therefore, we have to carefully select the value of time scale δ for the conventional model, since the estimates produced by the model depend on this value. This dependence can cause problems. For example, $k = 9.03079E+11$ when the value of time scale δ is equal to 1.

In this evaluation, we set δ equal to 0.1 for the conventional model, and δ equal to 1 for the proposed model. As is shown in Fig. 3, the first conventional model and the discrete model fit the actual data very well. The second conventional model is inferior to the other two.

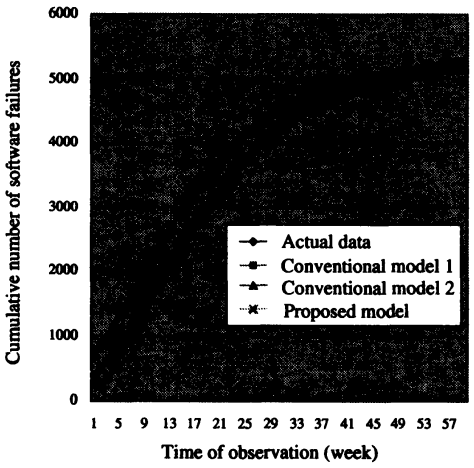

Figure 3: Comparison of both models with actual data.

However, the provision of accurate parameter estimates by a model is much more important early in the test phase than at the end of the test phase. Therefore, an important criterion for evaluating SRGMs is the accuracy of the parameter estimates they provide early in the testing phase. We compared both models on this criterion by estimating values for parameter k from increasing amount of data, starting with only the first small portion of data. As shown in Fig. 4, the values estimated by the proposed model were stabler than those estimated by both conventional models. The proposed model provides more accurate parameter values with the first small amount of data, so it provides a better way of estimating the number of potential software faults early in the testing.

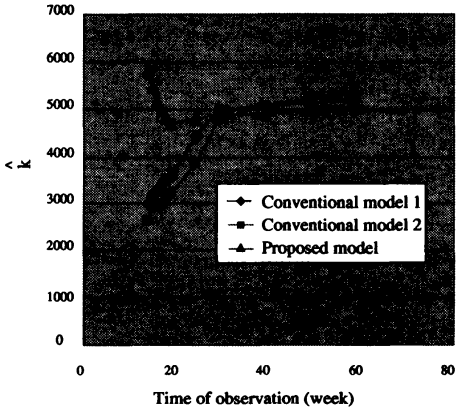

Figure 4: Estimate of parameter k.

4 Criterion for determining the absolute worth of a model

As has been shown in the previous sections, the proposed discrete models are capable of predicting total numbers of potential software failures on the basis of data gathered early in the test phase [28, 30].

In predicting total numbers of potential software failures, determining which model is the most appropriate model for use early in the testing phase is the next important and difficult task [7, 16].

We propose the following as a measure of the appropriateness of models [31]:

$$C = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \hat{X}_i}{X_i}\right)^2, \qquad (68)$$

where N denotes the number of available data points, X_i the actual data of the ith data point,

with $\hat{X}_i$ its value as estimated by an SRGM. Although error is usually evaluated as the mean squared error (MSE), the MSE is not fit for determining the appropriateness of models because it is significantly affected by the absolute values of the data. The proposed criterion, however, is not significantly affected by the absolute values of the data; rather, it is affected by the ratios between values of the data and estimates.

4.1 Evaluation on data sets that represent exact solutions

4.1.1 Data set A: The logistic equation

We analyzed the performance of the models thus for considered on the same four data sets as those in Sect. 2.

The result of the comparisons among the models is shown in Table 3, where C-l denotes the conventional logistic curve model 1 in Sect. 2.1.1, D-l denotes the discrete logistic curve model of Sect. 2.2, C-G denotes the conventional Gompertz curve model 1 of Sect. 3.1.1, and D-G denotes the discrete Gompertz curve model of Sect. 3.3. The discrete logistic curve model matched all the four sets of the data. This model reproduces the values of the parameters of the exact solution when the exact solution is used as the input data [30]. Thus, the values of criterion C in this case were all exactly zero. The conventional logistic curve model would be expected to provide a better fit in terms of criterion C because each data set of (A-i), ..., (A-iv) was composed of exact solutions of the logistic equation. However, for all data sets of (A-i), ..., (A-iv), the conventional logistic curve model provided a poorer fit than the conventional Gompertz curve model, as is shown in Table 3.

We then used each model to estimate k, the initial fault content. The results of comparison are shown in Table 4. The value of k as estimated by using the discrete logistic curve model matched the target value for all of the four data sets. The estimates of k provided by the conventional logistic curve model became more accurate as the number of available data points increased. Thus, in this case,

Table 3: Criterion C.

	C-l	D-l	C-G	D-G
A-i	7.4215	0	1.1420	0.15217
A-ii	15.697	0	0.99854	0.017104
A-iii	16.601	0	0.034532	0.0067477
A-iv	8.6398	0	0.012716	1.5459E-6

the estimate provided by using data set (A-i) gave a good approximation to the target value. The discrete and conventional Gompertz curve models, on the other hand, were much less accurate than the discrete and conventional logistic curve models. This was the case for all four data sets. Given the same target value of parameter k, the first several values of an exact solution to the Gompertz equation increase faster than those of the logistic equation. Hence, estimates by the Gompertz models were much larger than the target value.

Table 4: Estimated parameter k.

	C-l	D-l	C-G	D-G
A-i	99.23	100	88.11	200.0
A-ii	79.78	100	47.24	1.175E+6
A-iii	72.17	100	1.449E+7	9.702E+8
A-iv	60.17	100	1.198E+84	1.855E+184

4.1.2 Data set B: The Gompertz equation

In this case, we used the same data sets, (B-i), (B-ii), and (B-iii), as had been used in Sect. 3.

Comparative results for the models are given in Table 5. The discrete Gompertz curve model matched all three data sets. The discrete Gompertz curve model reproduces the values of the parameters of an exact solution when that exact solution provides the input data [28]. Thus, the corresponding values of criterion C were all exactly zero. Table 5 shows that the discrete and conventional logistic curve models provided a poorer fit in terms of criterion C

than did the discrete and conventional Gompertz curve models. This was the case for each data set in (B-i), ..., (B-iii). This result is reasonable because the data sets are from an exact solution to the Gompertz equation.

Table 5: Criterion C.

	C-l	D-l	C-G	D-G
B-i	0.3346	0.01454	2.195E-4	0
B-ii	1.141	0.01154	0.02100	0
B-iii	46.93	1.644E-32	0.2797	0

We used each model in estimating k. The comparative results are given in Table 6. The values of k estimated by the discrete Gompertz curve model from all the three data sets matched the target value. Estimates of k provided by the conventional Gompertz curve model became more accurate as the number of available data points increased. Thus, in this case, the estimate provided by using data set (B-i) gives a good approximation to the target value. However, the discrete and conventional logistic curve models were much less accurate than the discrete and conventional Gompertz curve models. This was the case for all the three data sets.

Table 6: Estimated parameter k.

	C-l	D-l	C-G	D-G
B-i	96.30	97.27	99.63	100
B-ii	31.44	55.36	78.16	100
B-iii	12.85	38.46	46.53	100

4.2 Evaluation on actual data sets

4.2.1 Data set C: Actual data set 1

We compared only the discrete logistic curve model and the discrete Gompertz curve model by using the same actual data set [30] as was used in Sect. 2, because both models yield accurate parameter estimates in the case of data that represent exact solutions, as was shown by the previous comparisons.

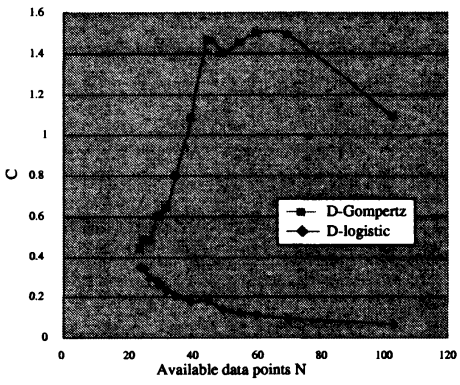

Figure 5: Criterion value vs. number of available data points.

We evaluated the parameter estimates for all of the data and for only that data available early in the test phase. We then used the estimated parameters to calculate values for criterion C. As is shown in Fig. 5, the discrete logistic curve model produced lower values for C than the discrete Gompertz curve model over the whole test phase.

We estimated k. The comparative results are shown in Fig. 6. The estimated values are normalized on the total number of actual software failures. The discrete logistic curve model provided more accurate parameter estimates. Moreover, this model provided accurate estimates throughout the range shown in the figure.

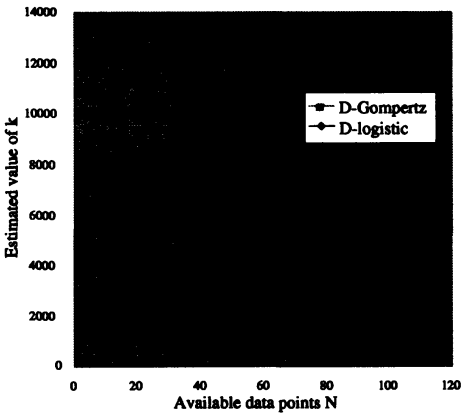

Figure 6: Estimates of parameter k.

4.2.2 Data set D: Actual data set 2

We used the same actual data set [14] as was used in Sect. 3. As was shown in Sect. 3, the discrete Gompertz curve model fit the actual data very well [28].

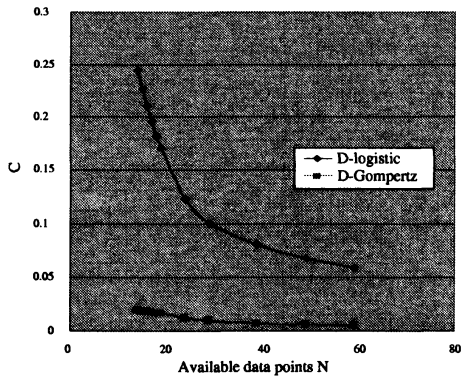

Figure 7: Criterion value vs. number of available data points.

We evaluated the parameter estimates for all the data and for only that the data available early in the test phase. We then used the estimated parameters to calculate values for criterion C. As is shown in Fig. 7, the discrete Gompertz curve model produced lower values for C than the discrete logistic curve model over the whole test phase.

We used each model to estimate k. The comparative results are shown in Fig. 8. The discrete Gompertz curve model provided the more accurate parameter estimates. Moreover, this model provided accurate parameter estimates from quite early in the test phase.

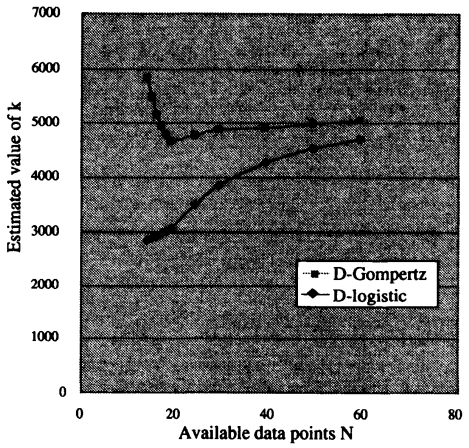

Figure 8: Estimates of parameter k.

5 Bass model

5.1 Bass model and conventional parameter estimations

Bass [1] suggested that the following differential equation can be used to represent the diffusion process:

$$\frac{dN(t)}{dt} = \left(p + \frac{q}{k}N(t)\right)(k - N(t)), \qquad (69)$$

where $N(t)$ is the cumulative number of adopters at a time t, k is the ceiling, p is the coefficient of innovation, and q is the coefficient of imitation. By introducing $F(t) = \frac{N(t)}{k}$, where $F(t)$ is the fraction of potential adopters who have adopted the product by time t, the Bass model can be restated as

$$\frac{dF(t)}{dt} = (p + qF(t))(1 - F(t)). \qquad (70)$$

If $N(0) = 0$, simply integrating both sides of equation (69) gives us the following distribution function to represent the time-dependent aspect of the diffusion process:

$$N(t) = k\left(\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}}\right). \qquad (71)$$

Equation (71) yields the S-shaped diffusion curve captured by the Bass model.

A number of procedures for estimating the parameters p, q, and k of the Bass model have been suggested. Mahajan *et al.* [11] compared the performance of four procedures—the ordinary least squares (OLS) [1], maximum likelihood estimation (MLE) [32], nonlinear least squares (NLS) [33], and algebraic estimation (AE) [10] procedures—on several sets of data. They concluded that NLS yielded better predictions as well as more valid estimates of standard error for the parameter estimates. On the other hand, the OLS is the easiest to implement. Therefore, we will look at the OLS and NLS procedures in detail in the following two sections.

5.1.1 Ordinary least squares procedure

The OLS procedure involves estimation of the parameters by taking the discrete or regression analogue of the differential equation (69). The regression equation is given as

$$X(i) = \alpha_1 + \alpha_2 N(t_{i-1}) + \alpha_3 N^2(t_{i-1}), \quad (72)$$

where

$$X(i) = N(t_i) - N(t_{i-1}), \quad (73)$$
$$\alpha_1 = pk, \quad (74)$$
$$\alpha_2 = q - p, \text{and} \quad (75)$$
$$\alpha_3 = -q/k. \quad (76)$$

Given regression coefficients[1] $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_3$, the estimates of parameters p, q, and k are easy to obtain:

$$\hat{p} = \frac{-\hat{\alpha}_2 + \sqrt{\hat{\alpha}_2{}^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2}, \quad (77)$$

$$\hat{q} = \frac{\hat{\alpha}_2 + \sqrt{\hat{\alpha}_2{}^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2}, \text{ and} \quad (78)$$

$$\hat{k} = \frac{-\hat{\alpha}_2 - \sqrt{\hat{\alpha}_2{}^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2\hat{\alpha}_3}. \quad (79)$$

The main advantage of the OLS estimation procedure is that it is easy to implement.

However, the OLS procedure has three shortcomings [32]. Firstly, as is clear from Eq. (72), in the presence of only a few data points and the likely multicolinearity of variables ($N(t_{i-1})$ and $N^2(t_{i-1})$), one may obtain parameter estimates that are unstable or possess wrong signs (examples [4, 32, 33]). Secondly, the standard errors of the estimates are not available since parameters p, q, and k are nonlinear functions of α_1, α_2, and α_3. The error term, however, does contain the net effect of all sources of error. Thirdly, the derivative of $N(t)$ which is obtained at t_{i-1} by the right-hand side of Eq. (73) will always be overestimated for time intervals before the point of inflection and underestimated after that. That is, a time-interval bias is present in the OLS approach since discrete time-series data are used to estimate a continuous-time model.

[1] $\hat{\alpha}_1 > 0$, $\hat{\alpha}_2 > 0$, and $\hat{\alpha}_3 < 0$ because $\hat{p}$, $\hat{q}$, and $\hat{k}$ are positive.

5.1.2 Nonlinear least squares estimation

The nonlinear least squares estimation procedure suggested by Srinivasan and Mason [33] was designed to overcome some of the shortcomings of MLE procedure [32], which itself was designed to overcome the shortcomings of the OLS procedure of Schmittlein and Mahajan [32]. From the cumulative distribution function given by

$$F(t) = \frac{1 - e^{-bt}}{1 + ae^{-bt}}, \quad (80)$$

Srinivasan and Mason suggest that parameter estimates $\hat{p}, \hat{q}$, and $\hat{k}$ can be obtained by using the following expression for the number of adopters $X(i)$ in the ith time interval (t_{i-1}, t_i):

$$X(i) = k\left(F(t_i) - F(t_{i-1})\right) + \mu_i, \quad (81)$$

where μ_i is an additive error term. Based on Eq. (81), parameters p, q, and k and their asymptotic standard errors can be directly estimated.

The NLS procedure overcomes the time-interval bias present in the OLS procedure. Furthermore, since the error term may be considered to represent the net effect of sampling errors, excluded variables (such as economic conditions and marketing mix variables), and mis-specification of the density function, the derived standard errors for the parameter estimates may be more realistic. However, since the NLS procedure employs various search routines in estimating the parameters, parameter estimates may sometimes be very slow to converge or may not converge, the final estimates may be sensitive to the initial values for p, q, and k, or the procedure may provide a non-global optimum.

5.2 Discrete Bass model

We propose a discrete Bass model, which is a form of a discrete Riccati equation [5]. The discrete Bass model enables us to forecast the diffusion innovation without using a continuous-time Bass model, because the discrete model has an exact solution.

The discrete Bass model is described as follows:

$$\frac{N_{n+1}-N_{n-1}}{2\delta} = p\left(k - \frac{N_{n+1}+N_{n-1}}{2}\right) + \frac{q}{k}\left(\frac{k}{2}(N_{n+1}+N_{n-1}) - N_{n+1}N_{n-1}\right). \quad (82)$$

The exact solution to equation (82) is written as

$$N_n = k\left(\frac{1-\left(\frac{1-\delta(q+p)}{1+\delta(q+p)}\right)^{\frac{n}{2}}}{1+\frac{q}{p}\left(\frac{1-\delta(q+p)}{1+\delta(q+p)}\right)^{\frac{n}{2}}}\right), \quad (83)$$

where $n = \frac{t}{\delta}$. This equation has also appeared in work on SRGM [37].

Applying OLS to the discrete Bass model is easy because the model is basically a time-discrete equation. The OLS procedure is the simplest method of parameter estimation for the discrete Bass model. In the continuous Bass model, the forward difference equation, which acts as a regression equation in the OLS procedure, is an approximation of the differential equation. However, in the discrete Bass model, the model itself is directly applied as the regression equation. Moreover, a solution of the discrete Bass model provides the same values as a solution of the continuous Bass model through the following equations:

$$p_d = \kappa p, \quad (84)$$

$$q_d = \kappa q, \quad (85)$$

$$\kappa = \frac{1}{\delta(p+q)}\left(\frac{1-\exp(-2(q+p))}{1+\exp(-2(q+p))}\right), \quad (86)$$

where p_d and q_d mean p and q in Eq. (83), respectively.

We propose two regression models. The first is the following equation:

$$S_n = 2(a+b(N_{n+1}+N_{n-1})+cN_{n+1}N_{n-1})+\epsilon(n), \quad (87)$$

where

$$S_n = N_{n+1} - N_{n-1}, \quad (88)$$

$$a = kp, \quad (89)$$

$$b = \frac{q-p}{2}, \quad (90)$$

$$c = -\frac{q}{k}, \quad (91)$$

$$\epsilon(n) : \text{error, and } E[\epsilon(n)] = 0. \quad (92)$$

Given regression coefficients[2] a, b, and c, parameter estimates $\hat{p}$, $\hat{q}$, and $\hat{k}$ are easily obtained as follows:

$$\hat{p} = -b + \sqrt{b^2 - ac}, \quad (93)$$

$$\hat{q} = b + \sqrt{b^2 - ac}, \text{ and} \quad (94)$$

$$\hat{k} = \frac{-b - \sqrt{b^2 - ac}}{c}. \quad (95)$$

The other regression model is the following equation:

$$M_n = A + BN_{n-1} + C(N_{n+1} - N_{n-1}) + \varepsilon(n), \quad (96)$$

where

$$M_n = N_{n+1}N_{n-1}, \quad (97)$$

$$A = \frac{k^2 p}{q}, \quad (98)$$

$$B = \frac{k(q-p)}{q}, \quad (99)$$

$$C = \frac{k(q-p-1)}{2q}, \quad (100)$$

$$\varepsilon(n) : \text{error, and } E[\varepsilon(n)] = 0. \quad (101)$$

Given regression coefficients[3] A, B, and C, parameter estimates $\hat{p}$, $\hat{q}$, and $\hat{k}$ are easily obtained as follows:

$$\hat{p} = \frac{-B + \sqrt{B^2 + 4A}}{2B - C}, \quad (102)$$

$$\hat{q} = \frac{B + \sqrt{B^2 + 4A}}{2B - C}, \text{ and} \quad (103)$$

$$\hat{k} = \frac{B + \sqrt{B^2 + 4A}}{2}. \quad (104)$$

These procedures have the advantage of simplicity, which is also provided by parameter estimation through the OLS procedure in the continuous Bass model.

Applying the NLS procedure to the discrete Bass model is also relatively easy, because the discrete Bass model has an exact solution (83). We propose two NLS procedures for the discrete Bass model. One of these provides estimated parameter $\hat{p}, \hat{q}$, and $\hat{k}$ by using the following expressions for the number of adopters X_n in the nth time interval:

$$X_n = N_{n+1} - N_{n-1} + \mu_n, \quad (105)$$

[2] $a > 0$, $b > 0$, and $c < 0$ because $\hat{p}$, $\hat{q}$, and $\hat{k}$ are positive.

[3] $A > 0$, $B > 0$, and $C < 0$ because $\hat{p}$, $\hat{q}$, and $\hat{k}$ are positive.

where μ_n is an additive error term.

The other NLS procedure for the discrete Bass model is the following equation:

$$Y_n = N_{n+1}N_n + \nu_n \qquad (106)$$

where Y_n is the ratio between the number of adopters at the nth time-step and that at the $(n+1)$st time-step.

These procedures have the advantage of allowing the direct estimation of the asymptotic standard errors of the parameters, as does the NLS procedure with the continuous Bass model. Moreover, since the error terms in these procedures may be considered to represent the net effect of sampling errors, excluded variables, and mis-specification of the density function, the derived standard errors for the parameter estimates may be as realistic as those of the NLS procedure for the continuous Bass model.

Either of the OLS procedures in the discrete Bass model overcomes the three shortcomings of the OLS procedure in the continuous Bass model: the time-interval bias, standard error, and multicolinearity.

When we use the discrete Bass model to avoid using the continuous model in forecasting the diffusion of innovation, there is no time-interval bias because the model is a discrete model. Furthermore, even if the discrete Bass model is regarded as only the procedure used to obtain the parameters for the continuous model, the OLS procedures do not suffer from a time-interval bias because a solution of the discrete Bass model gives the same values as a solution of the continuous Bass model, as was already stated in this section.

From Eq. (82), Eq. (87) is equivalent to Eq. (105), and Eq. (96) is equivalent to Eq. (106) under no constraints. Therefore, the same parameter estimation is done through both procedures in the discrete Bass model. This is a significant advantage of the discrete Bass model because we can get the global optimum through OLS, and then apply NLS to obtain the standard error. By using both procedures together, we overcome their shortcomings in separate application. That is, the standard error of the results of the OLS procedure is obtained through the NLS procedure. Equations (87) and (96) overcome the three shortcomings of NLS: that final parameter estimates are sensitive to the initial values of p, q, and k, that parameter estimates may sometimes be very slow to converge or may not converge at all, and that the optimum provided by the procedure may not be global.

Table 7 shows the condition number, the determinant of correlation matrix R, and the variance inflation factors (VIFs) for three procedures: the conventional OLS procedure (OLS), discrete analogue 1 of the OLS (87) (dOLS1), and discrete analog 2 of the OLS (96) (dOLS2), where we chose the exact solution ($p = 0.002$, $q = 1$, $m = 100$) to differential equation (69) as the data from every period from $t = 0$ to $t = 11$. The VIF in the conventional OLS row is the VIF of the variable $N(t_{i-1})$ in Eq. (72). From the definition of the VIF, the value of the VIF of the variable $N(t_{i-1})$ is the same as that of the VIF of the other variable, $N(t_i)^2$. The VIF in the dOLS1 row is the VIF of the variable $(N_{n+1} + N_{n-1})$; the VIF in the dOLS2 row is the VIF of the variables N_{n-1} in Table 7. dOLS2 excludes the problem of multicolinearity. Therefore, with this procedure, a wrong sign for a parameter suggests that the obtained data is not appropriate for the Bass model.

Table 7: Condition number, det R, and VIF.

Procedure	Condition number	det R	VIF
OLS	14.0111	0.01428	20.85
dOLS1	11.68	0.01914	12.68
dOLS2	3.548	0.2059	1.000

5.3 Parameter estimation

The accuracy of the parameter estimate provided by the conventional OLS procedure and the two OLS procedures in the discrete Bass model was compared. To compare the accuracy of the parameter estimates only, we chose data which satisfy the exact solution ($p = 0.002, q = 1, k = 100$) of differential equation (69) in every period from $t = 0$ to $t = 11$ (the same data as was used in the previous section). This data

has a point of inflection where $t^* = 6.2022$ and $N(t^*) = 49.9$. We analyzed three sets of data; data 1: the data up to the point just before the point of inflection ($t = 0, 1, \cdots, 6$), data 2: the data up to the point just after the point of inflection ($t = 0, 1, \cdots, 7$), and data 3: the data up to the ceiling ($t = 0, 1, \cdots, 11$).

The results of comparison of the conventional OLS, dOLS1 and dOLS2 procedures are given in Table 8. Both dOLS1 and dOLS2 provide accurate estimates. Since we used the exact solution to provide the data, an accurate procedure should reproduce the values of the parameters of the exact solution. Table 8 shows that both dOLS1 and dOLS2 reproduced k perfectly, even when the data did not include the point of inflection and there were fewer than eight data points.

Table 8: Estimated parameter k.

Data set	OLS	dOLS1	dOLS2
data 1	55.71	100	100
data 2	71.61	100	100
data 3	97.27	100	100

The accuracy of the conventional OLS procedure is poor despite the fact that the data is drawn from an exact solution of the differential equation. In particular, the conventional OLS procedure yields poor estimates of the parameters with data 1. This is consistent with the findings of Heeler and Hustad [4] and Srinivasan and Mason [33]. Through empirical studies, they found that stable and robust estimates of the parameters of the basic diffusion models cannot be obtained unless one uses at least eight data points, within which the point of inflection falls. The estimates of parameters with data 2 were also not accurate enough, even though data 2 satisfies the above condition.

Whenever a data set is a set from an exact solution of Eq. (69), the dOLS1 and dOLS2 procedures reproduce all values of the parameters, i.e., k, p, and q; theoretically, this is because the solution of Eq. (82) is the same as that of Eq. (69) through Eqs. (84), (85), and (86). This is independent of the number of data points and the values of the parameters. However, the conventional OLS procedure does not reproduce values of the parameters and depends on the number of data points, as shown in Table 8, because regression Eq. (72) does not have an exact solution and gives only an approximation of the Bass model.

We also evaluated the discrete Bass model on actual data. This data was the same as that used by Mahajan *et al.* [11], which was on the data diffusion of seven products: room air conditioners, color televisions, clothes dryers, ultrasound, mammography, foreign language, and accelerated program. These seven products represent a diverse set of innovations, and thus of sets of data, for all of which a minimum of eight annual data points covering the peak (point of inflection), is available. In addition, these products have been used extensively in the diffusion modeling literature to illustrate the application of alternative diffusion models or estimation procedures [1, 9, 32, 33].

To compare the predictive performance of the four estimation procedures, the OLS and the NLS procedure in the continuous Bass model and the two OLS procedures in the discrete Bass model, results related to a statistic of fit (MSE) are given in Table 9. The numbers (1, 2, $\cdots$, 7) in the left column represent, respectively, room air conditioners, color televisions, clothes dryers, ultrasound, mammography, foreign language, and accelerated program. The statistics of fit for dOLS2 is not directly comparable with those of the other estimation procedures, because the error term of dOLS2 is different from the error terms of the other estimation procedures. However, from Eqs. (87) and (96), the error term $\varepsilon(n)$ may be regarded as following

$$\varepsilon(n) = \frac{k}{q}\epsilon(n). \tag{107}$$

Therefore, we compared the fit statistics of dOLS2 with those of other procedures by using this equation.

Of the four procedures (the OLS, MLE, NLS, and AE procedures in the continuous Bass model), the NLS procedure provides the best fit to the data [11]. Mahajan *et al.* stated that, if we assume global optimum parameter

Table 9: Mean squared error.

	OLS	NLS	dOLS1	dOLS2
1	41,265	26,267	13,205	15,177
2	282,522	119,474	38,477	40,320
3	20,818	16,367	7,692	9,115
4	β	11.6	5.26	6.09
5	β	3.9	2.19	2.30
6	β	0.5	0.0949	0.0993
7	11.3	6.2	0.528	0.544

estimates, the NLS procedure should, by definition, provide the best fit in terms of the mean squared error [11]. However, a comparison of the statistics of fit in Table 9 indicates that both dOLS1 and dOLS2 provided a better fit to the data than did the OLS or NLS in terms of mean squared error. The fit statistic of dOLS1 was the best of all. A β in Table 9 indicates cases where the OLS procedure yielded an incorrect sign for the regression coefficient $\hat{\alpha_1}$ in the regression equation.

Table 10: Parameter estimates of k.

	OLS	NLS	dOLS1	dOLS2
1	17.1E6	18.7E6	18.0E6	17.1E6
2	35.5E6	. 39.7E6	39.1E6	38.4E6
3	15.3E6	16.5E6	16.19E6	15.3E6
4	β	167.4	187.2	180.2
5	β	111.4	122.1	121.2
6	β	37.6	40.1	39.6
7	63.6	64.4	65.5	65.1

Tables 10 and 11 show the parameters estimated by the OLS, NLS, dOLS1, and dOLS2 procedures. Again, β indicates where the OLS procedure yielded an incorrect sign for the regression coefficient $\hat{\alpha_1}$ in the regression equation. The results for the parameter estimates in Table 11 show that both dOLS1 and dOLS2 provided the wrong sign for the regression coefficient a in Eq. (87) and for the regression coefficient A in Eq. (96) for ultrasound, mammography, foreign language, and accelerated program. Both a in Eq. (87) and A in Eq. (96) are the regression coefficients of the constant term.

Table 11: Parameter estimates of p.

	OLS	NLS	dOLS1	dOLS2
1	0.0170	0.0094	0.0139	0.0107
2	0.0357	0.0185	0.02448	0.02194
3	0.0196	0.0136	0.01790	0.014322
4	β	0.0013	-0.01755	-0.02826
5	β	0.0004	-0.02501	-0.030308
6	β	0.0019	-0.0249	-0.02871
7	0.0120	0.0007	-0.01825	-0.0215363

A wrong sign in Table 11, however, does not indicate multicolinearity. Tables 12, 13, and 14, respectively, show the condition number, determinant of the correlation matrix, and variance inflation factors for each product. These tables show that there is no multicolinearity in dOLS2. Cases where the wrong signs were applied have smaller condition numbers, larger determinants of the correlation matrices, and smaller VIFs than the cases that had the right signs. Therefore, the wrong sign on a parameter suggests that the obtained data is not appropriate for the Bass model.

Table 12: Condition number.

	OLS	dOLS1	dOLS2
1	11.943	12.615	7.743
2	13.321	15.768	10.123
3	13.145	14.499	9.723
4	13.380	13.436	4.513
5	14.982	13.648	3.703
6	13.132	13.213	4.700
7	13.546	11.736	3.503

Table 13: Determinant of correlation matrix.

	OLS	dOLS1	dOLS2
1	0.01913	0.01614	0.03135
2	0.01453	0.009096	0.01152
3	0.01485	0.01138	0.01817
4	0.01565	0.01459	0.08556
5	0.01222	0.01383	0.1650
6	0.01658	0.01518	0.08084
7	0.01578	0.01973	0.1836

Table 14: Variance inflation factors.

	OLS	dOLS1	dOLS2
1	14.003	13.577	2.323
2	15.537	15.432	1.785
3	15.021	15.498	2.202
4	17.52	15.488	1.36
5	22.121	16.19	1.013
6	17.525	15.129	1.505
7	20.189	13.256	1.048

6 Discrete stochastic logistic curve model

As shown in the previous sections, the proposed discrete models yield accurate estimates of parameters, even with small amounts of input data. These models, however, are deterministic equations, so they do not yield the distribution of an estimate. In this section, we propose a discrete stochastic logistic equation that has an exact solution and then derive an SRGM from it, such that the distribution of an estimate is yielded along with the estimates themselves.

6.1 Discrete stochastic equation

We propose the following form of discrete stochastic logistic equation:

$$L_{n+1} - L_n = \delta \frac{A_{n+1}}{k} L_n (k - L_{n+1}), \quad (108)$$

where $\{A_n : n = 1, 2, \ldots\}$ is a sequence of independent and identically distributed (i.i.d.) random variables. Its exact solution is described by

$$L_n = \frac{k}{1 + m \prod_{j=0}^{n} \left(\frac{1}{1+\delta A_j} \right)}. \quad (109)$$

We suppose that $\{X_j : j = 1, 2, \ldots\}$ in Eq. (109)

$$X_j = \frac{1}{1 + \delta A_j} \quad (110)$$

has the i.i.d. power-function distribution. We consider the probability $P\{L_n > \underline{l}\}$ where

$$\underline{l} = \frac{k}{1 + m\underline{x}}. \quad (111)$$

Then, $P\{L_n > \underline{l}\}$ is described as follows,

$$\begin{aligned} & P\{L_n > \underline{l}\} \\ = \; & (\exp(\gamma \log \underline{x})) \sum_{j=0}^{n-1} \frac{(-\gamma \log \underline{x})^j}{j!}, \end{aligned} \quad (112)$$

Therefore, the proposed equation enables us to obtain a distribution for the estimate at a step n.

6.2 Distribution of actual data

The assumption of the power-function distribution stated in the previous section was evaluated on the same actual data as had been used in Sect. 2. We used the last value in the data series as the value of k. The distribution of X_j is shown in Fig. 9. Figure 9 indicates that X_j has the power-function distribution except at its tail, where the small amount of data leads to deviation from this distribution.

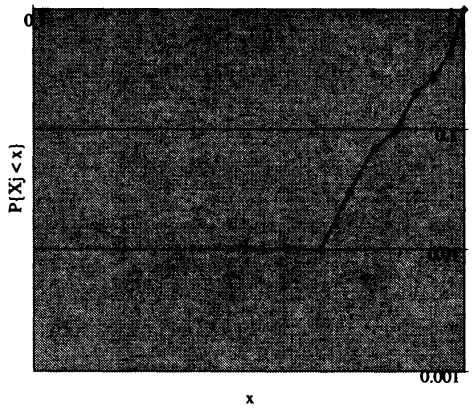

Figure 9: Distribution of the actual data

7 Conclusion

Through our discussions, we have shown that growth curve models based on integrable discrete equations provide characteristics which are beneficial in forecasting. We have applied these discrete models to software reliability models and a model of diffusion from marketing. The desirable properties of integrable discrete equations include their production of accurate estimates and accurate estimates of parameters early in the testing phase, and the fact that they avoid multicolinearity. These discrete models will be applicable to growth curve models in other fields, providing the same benefits as have been shown in this paper.

In Sect. 2, we proposed discrete logistic curve models, which are SRGMs that yield accurate parameter estimates from even small amounts of data. They are based on discrete forms of the logistic equation which were obtained by Morishita and Hirota. We constructed the SRGMs with no continuous components because the discrete equations have exact solutions. Morishita's and Hirota's models give the same parameter estimates. Although the conventional model uses a discrete equation as a regression equation, the model itself is a continuous time model, so it includes errors generated by discretization. However, our proposed models do not have this problem because they are themselves discrete models. We can thus analyze software reliability without having to use a continuous time model.

In Sect. 3, we proposed a discrete Gompertz equation that has an exact solution. The difference equation conserves the properties of the differential equation because it has an exact solution.

We also described an SRGM that yields accurate parameter estimates even from small amounts of data. It is based on the discrete Gompertz equation. Only regression analysis is need to estimate the parameters in the discrete model; this is also the case for the conventional model.

The proposed model enables us to accurately estimate parameters early in the testing phase, on the basis of actual data. In the conventional model, the parameter estimates change according to the number of data points. The discrete model provides parameter estimates that are stable against variation in the number of data points. This property is very important for an SRGM.

In predicting total numbers of potential software failures, it is also important, though difficult, to determine which model is the most appropriate for the early testing phase. In Sect. 4, we proposed a criterion for use in determining the most appropriate SRGM. This criterion, together with discrete SRGMs, determines the absolute worth of a model because the discrete SRGM perfectly reproduces the original parameters when the data used are from the exact solution of the equation with these parameters. The criterion is also applicable in determining the absolute worth of a discrete SRGM in operation on actual data sets. Therefore, the proposed criterion and the dis-

crete model enable us to identify the potential number of failures in an item of software early in the testing phase.

In Sect. 5, we described the discrete Bass model, a difference equation that has an exact solution. The discrete Bass model enables us to analyze diffusion processes. This discrete model has an exact solution, which is the same as the value of the corresponding solution of the continuous Bass model.

The two parameter estimation procedures in the discrete Bass model, the OLS and NLS procedures, give the same parameter estimates under no constraints. For actual data sets, both dOLS1 and the dOLS2 provide a better fit in terms of mean squared error, than the OLS or NLS procedure in the conventional Bass model. The parameter estimation procedures in the discrete Bass model are superior to the conventional procedures in terms of this criterion.

The parameter estimation procedures in the discrete Bass model also have certain broader advantages over those of the conventional Bass model. The OLS procedures of the discrete Bass model overcome the three shortcomings of the OLS procedure in the continuous Bass model: the time-interval bias, standard error, and multicolinearity. Although wrong signs on the parameters have been regarded as a problem caused by multicolinearity, we found that, in the case of the discrete method, wrong signs could be taken as indicators that the Bass model works poorly on the data.

In Sect. 6, a discrete stochastic logistic equation and an SRGM based on this equation were proposed. The equation has an exact solution and enables us to obtain a distribution of the estimate for any step, this is not possible with any discrete deterministic model.

In this paper, the discrete models have been shown to have six advantages over the conventional model. When the exact solution is used as the input data, the conventional model cannot reproduce the parameter estimates. It provides inaccurate parameter estimates when given data that do not encompass the inflection point. As has been done numerous times in the past, the accuracies were confirmed as being not too good, even with sufficient data points. However, the parameter estimation mechanisms in the discrete models reproduced the values of the parameters perfectly. Even if few data are given and the point of the inflection is not encompassed, the discrete models reproduce the values of the parameters either very accurately or perfectly. This is the first advantage.

The second advantage is that the discrete models are independent of time scale. We have to carefully choose the time scale for the conventional model because it must be used in the regression equation and the estimates depend on the choice of time scale. However, the time scale is not used in the discrete model's regression equation. The same parameter estimates are obtained whatever time scale we choose.

The third advantage is that the discrete models enable us to accurately estimate parameters from actual data gathered early in the testing phase. The parameter estimates of the conventional model vary with the number of data points. The discrete models provide stable values of parameter estimates for various numbers of data points. This property is very important for SRGMs.

The fourth advantage is that the criterion proposed for use with the discrete models allows us to determine which model is the most appropriate to use early in the testing phase. The proposed criterion and discrete models enable us to identify potential number of failures in software early in the testing phase.

The final advantage is that the discrete Bass model eliminates the multicolinearity of the conventional model.

The first and second advantages are provided by the exact solutions to the discrete equations. A given exact solution is equivalent to the exact solution of the differential equation. Although the other advantages are also provided by the exact solutions, it is difficult to see a direct relationship. Further studies investigation will be needed to determine these relationships.

References

[1] F.M. Bass: A new product growth model for consumer durables. *Management Sci-*

ence, **15** (1969) 215–227.

[2] W.D. Brooks and R.W. Montley: Analysis of discrete software reliability models, *Technical Report RADC-TR-80-84*, Rome Air Development Center, New York, 1980.

[3] R.L. Buchanan, R.C. Whiting, and W.C. Damert: When is simple good enough: a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves, *Food Microbiology*, **14** (1997) 313–326.

[4] R.M. Heeler and T.P. Hustad: Problems in predicting new product growth for consumer durables. *Management Science*, **26** (1980) 1007–1020.

[5] R. Hirota: Nonlinear partial difference equations. V. Nonlinear equations reducible to linear equations. *Journal of the Physical Society of Japan*, **46** (1979) 312–319.

[6] R. Hirota: Lecture on Difference Equations – From Continuous Values to Discrete Values (in Japanese), Saiensu-sha, Tokyo, 2000.

[7] A. Iannino, J.D. Musa, K. Okumoto, and B. Littlewood: Criteria for software reliability model comparisons, *IEEE Transactions on Software Engineering*, **10** (1984), 687–691.

[8] A. Kanno: Software Engineering (in Japanese). JUSE Press, Tokyo, 1979.

[9] S.B. Lawton and W.H. Lawton: An autocatalytic model for the diffusion of educational innovations. *Educational Administration Quarterly*, **15** (1979) 19–53.

[10] V. Mahajan and S. Sharma: A simple algebraic estimation procedure for innovation diffusion models of new product acceptance. *Technological Forecasting and Social Change*, **30** (1986) 331–346.

[11] V. Mahajan, C.H. Mason, and V. Srinivasan: An evaluation of estimation procedures for new product diffusion models. In V. Mahajan and Y. Wind (eds.): *Innovation Diffusion Models of New Product Acceptance* (Ballinger Cambridge, Massachusetts, 1986), 203–232.

[12] V. Mahajan, E. Muller, and F.M. Bass: New-product diffusion models. In J. Eliashberg and G.L. Lilien (eds.): *Handbooks in OR & MS 5: Marketing*, (Elsevier Science Publishers, 1993), 349–408.

[13] A. Messori: Survival curve fitting using the Gompertz function: a methodology for conducting cost-effectiveness analyses on mortality data, *Computer Methods and Programs in Biomedicine*, **52** (1997) 157–164.

[14] T. Mitsuhashi: A Method of Software Quality Evaluation (in Japanese). JUSE Press, Tokyo, 1981.

[15] F. Morishita: The fitting of the logistic equation to the rate of increase of population density. Research on Population Ecology, **VII** (1965), 52–55.

[16] J.D. Musa, A. Iannino, and K. Okumoto, Software Reliability: Measurement, prediction, application. McGraw-Hill, New York, 1987.

[17] A. Nagai, D. Takahashi, and T. Tokihiro: Soliton cellular automaton, Toda molecule equation and sorting algorithm, *Physical Letter*, **A255** (1999), 265–271.

[18] A. Nagai: Discrete integrable systems and convergence acceleration methods. Y. Nakamura (ed.): *Applied Integrable Systems* (in Japanese) (Shokabo, Tokyo, 2000), 225–260.

[19] Y. Nakamura: Integrable systems and algorithm. Y. Nakamura (ed.): *Applied Integrable Systems* (in Japanese) (Shokabo, Tokyo, 2000), 171–223.

[20] K. Nishinari and D. Takahashi: Analytical properties of ultradiscrete Burgers equation and rule-184 cellular automaton, *Journal of Physics*, **A31** (1998) 5439.

[21] M. Ohba, S. Yamada, K. Takeda, and S. Osaki: S-shaped software reliability growth curve: how good is it?, *Proceedings IEEE COMPSAC 82*, (IEEE Chicago, 1982) 38–44.

[22] M. Ohba: Inflection S-shaped software reliability groeth model, In S. Osaki and Y. Hatoyama (eds.): *Stochastic Models in Reliability Theory*, (Springer-Verlag, Berlin 1984), 144–165.

[23] K. Ohmori, and E. Shinohara: Predictive precision analysis of undiscovered errors, *Technical Report of IEICE* , **SSE98-190** (1999) 25–30.

[24] H. Ohmura: Topics in Forecasting (in Japanese). JUSE Press, Tokyo, 1993.

[25] M. Peleg: Modeling microbial populations with the original and modified versions of the continuous and discrete logistic equations, *Critical Reviews in Food Science and Nutrition*, **37** (1997) 471–490.

[26] H. Pasternak and B.A. Shalev: An algorithm to fit the Gompertz function to growth curves, *Computer Applications in the Biosciences*, **8** (1992) 239–241.

[27] T. Sakamaki: Software reliability – Software reliability forecast for quality management. *Technical Report of IEICE* , **R-81-8** (1981), 17–24.

[28] D. Satoh: A discrete Gompertz equation and a software reliability growth model, *IEICE Transactions*, **E83-D** (2000), 1508–1513.

[29] D. Satoh: A discrete Bass model and its parameter estimation, *Journal of the Operations Research Society of Japan*, **44-1** (2001), 1–18.

[30] D. Satoh and S. Yamada: Parameter estimation of discrete logistic curve models for software reliability assessment, *Japan Journal of Industrial and Applied Mathematics*, **19-1** (2002) 39–53.

[31] D. Satoh and S. Yamada: Discrete equations and software reliability growth models, *Proceedings of 12th International Symposium on Software Reliability Engineering*, (IEEE Computer Society, Hong Kong, 2001) 176-184.

[32] D. Schmittlein and V. Mahajan: Maximum likelihood estimation for an innovation diffusion model of new product acceptance. *Marketing Science*, **1** (1982) 57–78.

[33] V. Srinivasan and C.H. Mason: Nonlinear least squares estimation of new product diffusion models. *Marketing Science*, **5** (1986) 169–178.

[34] S. Tsujimoto, Y. Nakamura, M. Iwasaki: Discrete Lotka-Volterra system computes singular values, *Inverse Problems*, **17** (2001) 53–58.

[35] I. Walls and V.N. Scott: Validation of predictive mathematical models describing the growth of listeria monocytogenes, *Journal of Food Protection*, **60** (1997) 1142–1145.

[36] S. Yamada: Software Reliability Models – Fundamentals and Applications (in Japanese). JUSE Press, Tokyo, 1994.

[37] S. Yamada, S. Inoue, and D. Satoh: Statistical data analysis modeling based on difference equations for software reliability assessment (in Japanese), *Transactions of the Japan Society for Industrial and Applied Mathematics*, **12-2** (2002) 155–168.